



**QUEEN'S
UNIVERSITY
BELFAST**

Challenging the cancer molecular stratification dogma: Intratumoral heterogeneity undermines consensus molecular subtypes and potential diagnostic value in colorectal cancer

Dunne, P. D., McArt, D. G., Bradley, C. A., O'Reilly, P. G., Barrett, H. L., Cummins, R., ... Van Schaeybroeck, S. (2016). Challenging the cancer molecular stratification dogma: Intratumoral heterogeneity undermines consensus molecular subtypes and potential diagnostic value in colorectal cancer. *Clinical Cancer Research*. DOI: 10.1158/1078-0432.CCR-16-0032

Published in:
Clinical Cancer Research

Document Version:
Peer reviewed version

Queen's University Belfast - Research Portal:
[Link to publication record in Queen's University Belfast Research Portal](#)

Publisher rights
Copyright ©2016, American Association for Cancer Research.

General rights
Copyright for the publications made accessible via the Queen's University Belfast Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The Research Portal is Queen's institutional repository that provides access to Queen's research output. Every effort has been made to ensure that content in the Research Portal does not infringe any person's rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact openaccess@qub.ac.uk.

Challenging the cancer molecular stratification dogma: Intratumoral heterogeneity undermines consensus molecular subtypes and potential diagnostic value in colorectal cancer

Philip D Dunne ^{1*}, Darragh G McArt ^{1*}, Conor A Bradley ¹, Paul G O'Reilly ¹, Helen L Barrett ², Robert Cummins ², Anthony O'Grady ², Ken Arthur ¹, Maurice B Loughrey ^{1,3}, Wendy L Allen ¹, Simon S McDade ¹, David J Waugh ¹, Peter W Hamilton ¹, Dan B Longley ¹, Elaine W Kay ², Patrick G Johnston ¹, Mark Lawler ^{1#^}, Manuel Salto-Tellez ^{1#}, Sandra Van Schaeybroeck ^{1#}

¹ Centre for Cancer Research and Cell Biology, Queen's University Belfast, UK

² Department of Histopathology, Beaumont Hospital and Royal College of Surgeons in Ireland, Dublin, Ireland

³ Department of Histopathology, Royal Victoria Hospital, Belfast Health and Social Care Trust, Belfast, UK

*Joint First Authors

#Joint Senior Authors

^Corresponding Author

Conflict of Interest Disclosure: **PDD**: None; **DGMcA**: None; **CAB**: None; **PGOR**: None; **HLB**: None; **RC**: None; **AOG**: None; **KA**: None; **MBL**: None; **WLA**: None; **SSMcD**: None; **DJW**: is a Consultant for Almac Group and Almac Discovery; **PWH**: Founder, shareholder and Director in PathXL Ltd; **DBL**: None; **EWK**: None; **PGJ**: Previous Founder and Shareholder of Almac Diagnostics; CV6 Therapeutics: Expert Advisor and Shareholder; Chugai Pharmaceuticals: Consultant; **ML**: None; **MST**: None; **SVS**: None.

Running Title: Stromal-derived gene classifier underpins molecular subtype in colorectal cancer

Key Words: Colorectal cancer; gene expression profiling; molecular stratification; stromal derived gene classifier

Correspondence: Mark Lawler, Centre for Cancer Research and Cell Biology, Queen's University Belfast, Belfast, United Kingdom mark.lawler@qub.ac.uk

ABSTRACT

Purpose: A number of independent gene expression profiling studies have identified transcriptional subtypes in colorectal cancer (CRC) with potential diagnostic utility, culminating in publication of a CRC Consensus Molecular Subtype classification. The worst prognostic subtype has been defined by genes associated with stem-like biology. Recently, it has been shown that the majority of genes associated with this poor prognostic group are stromal-derived. We investigated the potential for tumor misclassification into multiple diagnostic subgroups based on tumoral region sampled.

Experimental design: We performed multi-region tissue RNA extraction/transcriptomic analysis using Colorectal Specific Arrays on invasive front, central tumor and lymph node regions selected from tissue samples from 25 CRC patients.

Results: We identified a consensus 30 gene list which represents the intratumoral heterogeneity within a cohort of primary CRC tumors. Using a series of online datasets, we showed that this gene list displays prognostic potential (HR=2.914 (CI 0.9286-9.162) in stage II/III CRC patients, but in addition we demonstrated that these genes are stromal derived, challenging the assumption that poor prognosis tumors with stem-like biology have undergone a widespread Epithelial Mesenchymal Transition (EMT). Most importantly, we showed that patients can be simultaneously classified into multiple diagnostically relevant subgroups based purely on the tumoral region analysed.

Conclusion: Gene expression profiles derived from the non-malignant stromal region can influence assignment of CRC transcriptional subtypes, questioning the

current molecular classification dogma and highlighting the need to consider pathology sampling region and degree of stromal infiltration when employing transcription-based classifiers to underpin clinical decision-making in CRC.

Translational relevance: Recent molecular profiling studies in CRC have identified distinct subgroups with significant prognostic value. These studies have proposed the use of transcriptional-based companion diagnostic tests, based on these prognostic subgroups. In stage II/III CRC, such diagnostic tests are urgently needed to inform clinical decision-making in regard to administration of adjuvant therapy following surgery. The identification of a poor prognostic group has been associated with EMT- and Stem-like traits, although detailed analysis of these tumors has revealed that the stromal-rich architecture of the tumor microenvironment informs their biology. In this study, we highlight the contrasting gene expression at multiple regions of CRC primary tumors due to intratumoral heterogeneity. This differential biology itself displays prognostic value and overlaps with the proposed diagnostic classifier signatures in CRC. Further to this, we highlight the potential for individual patient misclassification due to this regional heterogeneity using the recently proposed molecular classifiers.

INTRODUCTION

Colorectal cancer (CRC) is the second leading cause of cancer-related deaths in the Western world. To date, the most powerful prognostic classification tool in CRC still relies on the Tumor Node Metastasis (TNM) staging system. While histopathology is able to identify a number of factors which provide additional prognostic value and capture biological aspects of important disease processes, (such as tumor morphology, grade and stage, lymphovascular invasion and peritumoral inflammation), it gives no insight into the distinct biology that underpins CRC and determines disease outcome.

We and others have employed transcriptional profiling of clinically-annotated tumor samples to gain molecular insights into CRC biology and thereby develop new patient stratification tools and identify novel therapeutic approaches (1-6). Using unsupervised classification of transcription data, independent groups have reported 3-6 molecular subgroups within CRC (2, 4, 7). Recent subtype concordance analysis by the CRC Subtyping Consortium (CRCSC) has yielded a consensus of 4 transcriptionally-driven CRC molecular subgroups (Consensus Molecular Subtype (CMS 1-4)) with the following distinguishing features: CMS1: Microsatellite instable(MI)/immune (frequency(f) =14%); CMS2: Canonical (f=37%); CMS3: Metabolic (f=13%) and CMS4: Mesenchymal (f=23%) (8, 9). Of these subgroups, the mesenchymal subtype (CMS4) has been found to be significantly associated with the worst prognosis in all studies. Additionally, this gene signature has also been detected in serrated precursor lesions, suggesting that aggressive tumor traits may be “imprinted” in dysplastic tissue at the very earliest stage.

Many of the studies highlighted above (2, 4, 7) suggested that the identification of the CMS4 subtype was evidence of epithelial tumor cells undergoing changes associated with an extensive epithelial-mesenchymal transition (EMT), resulting in increased invasion and the acquisition of stem cell properties that are necessary for the establishment of metastases (10). Although the findings of these studies were independently validated, the cells-of-origin of individual transcriptional subtypes were not defined. More recently, two studies have indicated that mesenchymal tumors have an enriched stromal component, suggesting that the origin of the CMS4 signature may be dependent on the transcriptional profile originating from the stromal rather than the tumor compartment (11, 12). In this paper, we build on these recent studies and provide important insights into the role of the colorectal tumor stroma, which may have significant implications for transcription-based classifiers as precision medicine tools for patient stratification.

Materials and Methods

Tissue samples

Upon receipt in the histopathology laboratory, fresh colon resection specimens (resected for tumor) were examined by a pathologist. Regions of interest (deep and superficial tumor) were identified using standard approaches and described for subsequent sampling. Specimens were then fixed in 10% neutral buffered formalin, sampled and processed as per routine histology. Subsequent examination of 3 μ m H&E stained sections confirmed the presence of the identified regions of interest and examination of further material from the same resection identified lymph nodes positive for tumor metastasis. The areas to be sampled were marked on the slides by a pathologist and these were then used as a guiding template for macro-dissection of 10 μ m unstained tissue sections cut from the same block. Tissue of interest was scraped off the slide using a scalpel and a small amount of ethanol and placed in a labelled micro-centrifuge tube.

Sample analysis

We selected samples from 25 patients (for clinical details see Supplemental Table 1) where we had sufficient high quality material to extract RNA from regions of interest. This resulted in 75 transcription profiles, 25 patients' samples at 3 regions per sample. These 75 samples were analyzed on the CRC disease specific array (DSA) platform (Almac). Initial QC analysis of the microarray data found that 2 samples were outliers and were subsequently removed from further downstream analyses, resulting in 73 transcriptional profiles; 24 invasive front, 24 central and 25 lymph node deposits.

Data analysis and classifier generation

DSA arrays were read into the R statistical environment (version 3.1), background corrected and normalized using the *justRMA* function in the *affy* package. We initially assessed the profiles for dissimilarities by bootstrap hierarchical clustering to assess confidence in the groups and their distances by *pvclust*, allowing also for various non-specific gene filtering levels by the package *genefilter*. From here, the *limma* package was used to assess MDS plots as well as to perform differential expression analysis between contrasts and generation of the overlap by Venn diagram. In order to create a succinct list of genes that could best classify the central region against the invasive front, we adopted a machine learning method, *pamr*. Groups were stratified randomly into a three quarter training set and one quarter test set. We used *genefilter*'s 'varFilter' method to reduce the complexity of the features, which we also repeated for the *pamr* process at increasing stringencies to increase confidence in the list generated. We performed *pamr* analysis using a ten-fold cross-validation and chose our threshold to select the lowest misclassification rate ($t = 2.0$, misclassification = 5/36 and an overall error rate = 0.138). Using our initial classification run, *pamr* proffered a list of 58 features equating to 30 annotated genes. This annotated gene list formed the benchmark for dataset stratification in other independent datasets for the rest of this study. Hierarchical clustering performed in R was done using *heatmap.plus* using the Ward clustering method with Euclidean distance and with ordination of the data performed using correspondence analysis with the package *made4* and visualized using the *plotarrays* parameter with differing *graph* arguments.

Independent datasets

Affymetrix gene expression profiles from independent CRC datasets were downloaded from NCBI Gene Expression Omnibus (GEO) (<http://www.ncbi.nlm.nih.gov/geo/>) under accession numbers GSE14333, GSE35602, GSE33113, GSE39396 and GSE39582. In addition, the independent mouse dataset was accessed under GSE6383. Partek Genomics Suite was used on these series matrix files for independent dataset analysis. For the purpose of clustering, the data matrices were standardized to the median value of probe sets expression. Standardization of the data allows for the comparison of expression levels for different probe sets', which may not necessarily be on the same scale or at the same intensity levels. Following standardization, 2-dimensional hierarchical clustering was performed using the Ward clustering method with Euclidean distance.

Statistics

Positive prediction value was calculated as $PPV = \text{number of true positives} / \text{number of positive calls}$. Negative prediction value was calculated as $NPV = \text{number of true negatives} / \text{number of negative calls}$. Sensitivity = $\text{number of true positives} / \text{number of true positives} + \text{number of false negatives}$. Specificity = $\text{number of true negatives} / \text{number of true negatives} + \text{number of false positives}$.

RESULTS

Identification of region-specific transcriptomic profiles:

In order to assess tumor heterogeneity at the level of transcription, we performed gene expression profiling on a cohort of patients (n=25, Supplementary Table 1) using tumor samples taken from multiple regions of the primary tumor, in addition to tumor tissue which had metastasized to regional lymph nodes (Supplementary Figure 1). The resulting transcriptomic data were analyzed using *limma* to identify differentially expressed genes. A supervised learning approach, *pamr*, was employed to extract a succinct list of genes that could highlight the contrasts between samples from different regions, which could then be visualized using a combination of a heatmap and hierarchical clustering.

Determination of the number of genes which were differentially expressed (DE) between each region indicated that the transcriptome of the central tumor (CT) was more closely related to that of the invasive front (IF) than to that of the lymph node (LN) tumor tissue (Figure 1A and Supplementary Tables 2-4). However, both *limma* and *pamr* analyses revealed distinct differences between gene expression in the CT and IF, with a consistent upregulation of known drivers of disease progression, such as IGF1 (2.3–5-fold) and SFRP2 (2.6–4-fold) at the IF region, while upregulation of CXCL5 (3.3–8-fold), IL8 (3.6–4.4-fold) and MMP1 (2.6–7.9-fold) were consistently found in the CT region (Figure 1B, Supplementary Table 2).

Using *pamr* classification, we extracted a concise gene list, comprising 30 annotated genes, that could identify and characterize the transcriptome associated with a minimum overall error (error rate = 0.138, Table 1). As expected, this gene list

contained a number of genes already reported to be involved in cancer progression, many of which were also represented in the DE gene list (Supplementary Table 2), such as IGF1, TGFB3, IL8, IL1B and WNT5A. We validated the gene list by applying it to the full cohort of IF and CT samples (Figure 1C) using hierarchical clustering to visualize ranges of expression. The region-of-origin was correctly identified in 44 out of 48 samples, giving this classifier a positive prediction value (PPV) for membership of IF equal to 91.7% and a corresponding negative prediction value (NPV) for membership of CT equal to 91.7%.

Specific gene signatures identify tissue regions of origin

To test the ability of stromal gene expression to classify tumor samples, we utilized a list of 213 genes, which are specifically expressed in cancer-associated fibroblasts (CAFs) (n=131), leukocytes (n=47) or endothelial cells (n=35) (11). Using this gene list, the samples from LN, CT and IF could be separated into 3 biological groups corresponding to their site of origin, using correspondence analysis of the top quartile most variant probes (Figure 2). Whereas Isella *et al.* (11) and Calon *et al.* (12) have shown that these 213 non-epithelial genes can be used to identify poor prognostic patients, we now demonstrate that the same genes can be used to identify the *site* of the tumor from which the sample has been taken. Moreover, the 30 gene list (which we now define as a stromal classifier) was also able to correctly distinguish between tumor and stroma tissues in a cohort of 13 CRC tumor samples (GSE35620), each of which was micro-dissected into these two components, with no misclassification (Figure 3A, 3B). As a further validation, we used a tissue-specific dataset (GSE6383), which identified genes expressed in the mesenchymal and epithelial compartments of the perinatal mouse intestine (13). This analysis revealed

that the genes which we found to be overexpressed in IF compared to CT are highly expressed in the stromal mesenchyme fraction of the developing mouse intestine compared to the epithelium, and can stratify samples into regions of origin with no misclassification (Figure 3C). Collectively, these findings indicate that our 30 gene stromal classifier is an accurate identifier of tissue origin.

To further delineate the source of the identified transcriptional signals to a precise cell-of-origin, we utilized microarray profiles derived from primary CRC tumors which have undergone disassociation and sorting in order to isolate cells into epithelial, endothelial, fibroblast and leukocytes compartments (GSE39396). Hierarchical clustering of our classifier using these cell sorted profiles indicated the upregulation of each transcript signal specifically in the fibroblast component, compared to the remaining tumor cell types (Figure 4A). Following median normalization for each individual gene across all four cell types, we again confirmed that the upregulation of the transcription levels for our classifier occurs specifically within the fibroblasts, compared to all other cell types (Figure 4B).

The 30 Gene Stromal Classifier overlaps with the CMS4 subtype

We next interrogated an independent stage II/III colon cancer dataset (2). Using a semi-supervised clustering approach, we identified a subset of tumors accounting for 20% of the overall cohort that were positive for the 30 gene Stromal Classifier. Of these tumors, almost 90% are associated with the previously reported stem-like (CMS4) poor prognostic group (2) (Figure 5A). These findings show that the genes which are differentially expressed between the central and invasive front regions of a primary tumor can identify patients exhibiting a poor prognostic “stem cell-like”

mesenchymal biology. Additionally, in 2 further independent datasets GSE14333 (a mixed Stage II/III dataset (14)) and GSE33113 (an untreated stage II dataset (7)), the 30 gene Stromal Classifier consistently identified patients with a poorer prognosis, although these associations just failed to reach significance, most likely due to limited sample numbers (Figure 5B).

To assess the extent that patient classification can be dictated by the tumor region sampled, we performed a reciprocal *in silico* analysis, in which the 48 samples representing IF and CT were classified using the genes previously published to classify the stem-like mesenchymal subtype (2). This analysis revealed 2 separate clusters, both highly enriched for samples based on the region-of-origin, with 75% of profiles originating from the IF being classified as stem cell-like (Figure 5C), compared to 33% for CT.

Discussion

Region-specific transcriptional profiling allows precise dissection of the role of different tissue compartments in CRC disease biology, while also providing prognostic information on individual patients. Our results from 2 independent datasets GSE35620 (Figure 3A, 3B) and GSE6383 (Figure 3C) are in line with the recent stromal studies reported (11, 12) and further confirm that the key genes involved in classification of patient samples are derived from a stromal origin. These data collectively challenge the general assumption that tumors classified as mesenchymal, or in the CMS4 subgroup, have undergone widespread EMT, resulting in lower levels of epithelial associated traits in the tumor cell compartment, when actually these tumors have a higher component of stromal (particularly fibroblast), infiltration. Although immunohistochemistry-based analysis of CRC tumors has shown that neoplastic epithelial cells expressing stem-like properties have a poor prognosis (15-17), our findings emphasize that the CMS4 subgroup represents tumors with higher transcription levels of mesenchymal-associated genes, which can be attributed to their overall stromal-rich, and in particular fibroblast, architecture.

While our data strongly suggest that the overall proportion of EMT in CMS4 tumors is not as extensive as first proposed, we did observe some evidence of localized loss of E-cadherin staining in a small proportion of budding cells (data not shown). Precise examination of these individual tumor buds may reveal the biology driving the invasive nature of individual cells compared to the overall tumor mass. Stromal-epithelial interactions have a clear underlying role in the development of tissue and in the shaping of the microenvironment in both normal and malignant conditions. The

configuration of the tumor border region itself can also give insight into the prognosis of a patient, with a high degree of stromal infiltration into a tumor being a poor prognostic indicator (18, 19).

While the finding that high levels of stromal infiltration are associated with poor prognosis is not new (11, 12, 18, 19), the present study precisely delineates the cell-of-origin to the fibroblast compartment, with primary results validated in a comprehensive series of online datasets. Data presented here, in addition to these referenced studies (12, 18-20) highlight that expression of a small number of genes in fibroblasts associated with the tumor, in combination with the overall fibroblast content, is sufficient to identify poor prognostic patients in stage II CRC. It is unsurprising therefore that the prognostic value of the CMS4 subtype, and indeed other current transcriptional-based prognostic classifiers, is highly dependent on stromal-derived gene signatures and that our 30 gene Stromal Classifier also identifies a group of patients with a worse prognosis. A combined classification system, utilizing clinical, morphological and molecular features, has been previously proposed and shown to provide a confident prediction of patient outcome (21). Multi-region sequencing and mutational analysis has highlighted diversity in renal (22) and recently in CRC (23) patient samples, but our data shows that patient classification in CRC utilizing transcriptional profiling is associated as much with the region from where the sample is derived from, as with the underlying tumor biology.

Implications of transcriptomic heterogeneity within CRC

Data presented here shows for the first time that the recently published CRC CMS classification system (9) is dependent not just on the overall driving biology of a

tumor, but also on the region of sampling. We have shown that each individual primary tumor has the potential to be classified into a poor prognostic mesenchymal subtype if the tissue employed in the analysis is taken from the invasive front of the tumor, regardless of the inherent CMS subtype of the tumor overall. These findings may have important implications for the use of molecular classification approaches in selecting patients for therapeutic intervention in biomarker-driven stratified clinical trials. Given the lack of standardized procedures employed in tissue sampling for these transcriptional profiling approaches, our data provide cautionary evidence that a generic transcriptional profiling strategy, which does not consider the role of specific site of tissue origin within the tumor, could result in patient misclassification into inappropriate subgroups, leading to inaccurate prognostication and treatment that yields no benefit or may even potentially harm patients.

A recent precision medicine-based clinical trial, the WINTHER study (24, 25) (an output of the Worldwide Innovative Networking in personalized cancer medicine (WIN) Consortium), aimed at selecting optimal innovative drug treatment strategies for patients with advanced metastatic tumors, has highlighted the utility of sampling matched normal and tumor biopsies from the same patient in order to identify novel therapeutic strategies based on differential gene expression in the neoplastic tissue compared to the normal sample. There are a number of clear similarities between the WINTHER study approach and the findings presented here, where the non-tumor component can inform prognosis, but while there seems to be clear rationale for this approach aimed at personalized drug selection, there may be only limited benefit derived by subtracting the results of adjacent normal or stromal tissue in order to correctly classify the overall tumor molecular subtype. As the defined CMS of CRC

are based on the presence and extent of infiltrating stroma, the subtraction of non-tumor cell signaling would likely result in further misclassification, similar to what *Isella et al.* (11) reported following molecular subtyping of PDX implanted CRC tumor tissue.. Furthermore, our data and those of *Isella et al* (11) and *Calon et al* (12) also raise the important question as to which aspect of this stromal-tumor interface dictates aggressive tumor biology, and in turn, response to therapy.

The primary tumor samples used in this study are predominantly from patients with non-metastatic disease (92% stage II/III – Supplementary Table 1) and as such represent samples with limited molecular heterogeneity compared to pretreated metastatic solid tumor samples or samples from advanced leukemia patients (e.g. blast crisis Chronic Myeloid Leukemia (CML)), both of which can introduce further complicating issues when assessing prognosis and treatment response using molecular-based precision medicine approaches (26). While there is, as yet, no defined metastatic-specific CMS typing approach which captures transcriptional changes associated with liver and lung tissue infiltration, there will inevitably be variations in the stromal content of metastatic lesions in both the liver and lung similar to what we have presented here in the primary setting. The prognosis of tumors with high metastatic site-specific stromal infiltration may not equate to that of high levels of fibroblast infiltration in primary CRC tumors, and in a similar scenario to blast crisis CML, it is clear that putative treatment regimes for molecularly heterogeneous metastatic neoplastic lesions from solid tumors require very careful consideration in order to achieve therapeutic success (26).

In conclusion, the data we have presented here, alongside other recent publications (11, 12), question the contribution of tumor epithelial cells to CRC gene expression signatures and emphasize the need to value the role of stromal-derived signatures in modulating CRC biology. Furthermore, they also provoke a cautionary re-evaluation of the thesis of several recent publications (2-4) (including the CRCSC CMS paper (9)) that molecular classification, irrespective of tumor tissue site of origin, can be reliably employed to underpin treatment algorithms in CRC. We propose that a morpho-molecular strategy (27), integrating tissue morphology/origin, molecular evaluation (e.g. gene expression profiling, next generation sequencing, tissue microarray analyses) and clinical phenotype in a common integromics framework (28), is necessary for precise, robust and clinically-relevant dissection of tumor biology that will inform the next generation of investigator-led biomarker-driven interventional trials in this common disease.

Figure legends

Figure 1: Differential analysis and classifier generation.

A. Venn diagram, produced using *limma*, of numbers of probesets which are differentially expressed between each region (IF = Invasive Front; CT = Central Tumor; LN = Lymph Node). The total number of differentially expressed genes increases as spatial distance from the central tumor increases. The number of probes which do not appear in any of the three Venn sets is displayed in the bottom right. This indicates that there are very small but robust differences in the transcriptome of the CT compared to the IF **B.** Volcano plot depicting highest ranked differentially expressed genes between the central and invasive front regions. Gene symbols corresponding to each individual probeset identified are depicted. **C.** Heatmap of central and invasive front samples clustered by expression values of the 30 genes in the supervised invasive classifier. Overlay bar depicts the region from which the sample originated, with black representing the central tumor and grey representing invasive front samples, confirming the robustness of the 30 genes used in classifying samples based on region-of-origin.

Figure 2: Ordination of samples by region-of-origin based on previously published stromal-specific gene signature

Correspondence analysis depicting a strong separation and grouping of the CT, IF and LN samples according to their region-of-origin using the most variant probes from the Isella *et al.* 213 stromal-specific gene signature following cross annotation (top quartile). Left pane depicts the projection of individual samples with the right pane depicting the grouping with ellipse and star arguments. These findings highlight

the role of the non-tumoral transcriptome on the intratumoral heterogeneity of primary CRC tumor tissue analysis.

Figure 3: Stromal derived gene transcripts underpin sample classification according to region-of-origin

A. Heatmap based on expression values of the genes from our supervised stromal classifier which are highly expressed at the invasive front can stratify samples according to tissue type in GSE35620, which contains transcriptional profiles of microdissected primary CRC tumors purified into epithelial and stromal compartments. **B.** PCA plot from sample distribution based on clustering in A, again validating the utility of our classifier in identifying tissue type. **C.** Heatmap based on expression values of genes from the stromal classifier which are highly expressed at the invasive front, using transcriptional profiles from GSE6383, which contains mouse intestinal tissue separated into epithelial and mesenchymal compartments. These findings highlight the stromal-specific source of the gene transcripts identified in our classifier.

Figure 4: Specific cell of origin of transcriptional classifier

A. Heatmap expression values of stromal classifier in GSE39396 of Fluorescence Activated Cell Sorting (FACS) selected cells into specific endothelial [CD45(+), EPCAM(-), CD31(-), FAP(-)], epithelial [CD45(-) EPCAM(+), CD31(-), FAP(-)], leukocyte [CD45(-), EPCAM(-), CD31(+), FAP(-)] and fibroblast [CD45(-), EPCAM(-), CD31(-), FAP(+)] populations which suggests upregulation of these transcripts in the cancer-associated fibroblast compartments compared to the other compartments of the tumor microenvironment. **B.** These findings were confirmed by median

normalizing and plotting each gene using dot-plots according to specific cell type. Median value for overall gene expression in each cell type is indicated. Error bars indicate standard deviation.

Figure 5: Prognostic value and patient stratification using transcriptional profiling is based on stromal derived genes.

A. Identification of an invasive-like cluster, accounting for 20% of the patient cohort, using our supervised stromal classifier and expression data from a large well characterized stage II/III CRC dataset; GSE39582. This cluster strongly identifies the previously characterized stem-like C4 (CMS4) poor prognostic group with a sensitivity of 89% and specificity of 87%. **B.** Euclidean clustering and survival analysis using our stromal classifier in GSE14333 (top panels) and GSE33113 (bottom panels). The stromal classifier isolates a subgroup of patients accounting for between 21-26% of each cohort, with a trend towards a worse prognosis compared to remaining patients (HR=1.942 in GSE14333 and HR=2.914 in AMC-AJCCII-90). **C.** Heatmap depicting Euclidean clustering of central and invasive front samples using previously reported stem-like patient classifier (2). This poor prognostic classifier defines two distinct groups, based on clustering distance, which are highly stratified based on region-of-sampling using the most variant probes (top quartile).

Table 1: 30 gene classifier of tumor region.

Top 30 ranked genes across returned from our supervised analysis giving the lowest misclassification rate, which can identify the region-of-origin when using multiple samples from the primary tumor with a PPV and NPV of 91.7%. The top 20 genes (in red) are over-expressed in the invasive front samples compared to the central tumor, whereas the bottom 10 genes (in green) are over-expressed in the central tumor samples compared to the invasive front. This machine learning derived 30 gene classifier can identify tumor tissue based on region-of-sampling as shown in Figure 1C.

Authors Contributions

Conception and Design: PDD, DGMcA, ML, MST, SVS

Development of Methodology: PDD, DGMcA

Acquisition of Data: HLB, RC, KA, AOG, MBL, EWK

Analysis and Interpretation of Data: PDD, DGMcA, CAB, PGOR, WLA, PWH, SSMcD

Writing and Reviewing Manuscript: PDD, DGMcA, MBL, DJW, EWK, SSMcD, DBL, PGJ, ML, MST, SVS

Administrative, technical or material support: KA

Study Supervision: PGJ, ML, MST, SVS

Grant Support

This work was supported by a Cancer Research UK (CRUK) Programme Grant and a joint Medical Research Council UK (MRC) and CRUK Programme Grant (S:CORT)

References

1. Kennedy RD, Bylesjo M, Kerr P, Davison T, Black JM, Kay EW, et al. Development and independent validation of a prognostic assay for stage II colon cancer using formalin-fixed paraffin-embedded tissue. *J Clin Oncol*. 2011 Dec 10;29(35):4620-6.
2. Marisa L, de Reynies A, Duval A, Selves J, Gaub MP, Vescovo L, et al. Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. *PLoS Med*. 2013;10(5):e1001453.
3. De Sousa E Melo F, Wang X, Jansen M, Fessler E, Trinh A, de Rooij LP, et al. Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions. *Nat Med*. 2013 May;19(5):614-8.
4. Sadanandam A, Lyssiotis CA, Homiczko K, Collisson EA, Gibb WJ, Wullschlegel S, et al. A colorectal cancer classification system that associates cellular phenotype and responses to therapy. *Nat Med*. 2013 May;19(5):619-25.
5. Roepman P, Schlicker A, Tabernero J, Majewski I, Tian S, Moreno V, et al. Colorectal cancer intrinsic subtypes predict chemotherapy benefit, deficient mismatch repair and epithelial-to-mesenchymal transition. *Int J Cancer*. 2014 Feb 1;134(3):552-62.
6. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*. 2012 Jul 18;487(7407):330-7.
7. de Sousa E Melo F, Colak S, Buikhuisen J, Koster J, Cameron K, de Jong JH, et al. Methylation of cancer-stem-cell-associated Wnt target genes predicts poor prognosis in colorectal cancer patients. *Cell Stem Cell*. 2011 Nov 4;9(5):476-85.
8. Dienstmann R, Guinney J, Delorenzi M, De Reynies A, Roepman P, Sadanandam A, et al. Colorectal Cancer Subtyping Consortium (CRCSC) identification of a consensus of molecular subtypes. *ASCO Meeting Abstracts*. 2014 June 11;32(15_suppl):3511.
9. Guinney J, Dienstmann R, Wang X, de Reynies A, Schlicker A, Soneson C, et al. The consensus molecular subtypes of colorectal cancer. *Nat Med*. 2015 Oct 12.
10. Loboda A, Nebozhyn MV, Watters JW, Buser CA, Shaw PM, Huang PS, et al. EMT is the dominant program in human colon cancer. *BMC Med Genomics*. 2011 Jan 20;4:9,8794-4-9.
11. Isella C, Terrasi A, Bellomo SE, Petti C, Galatola G, Muratore A, et al. Stromal contribution to the colorectal cancer transcriptome. *Nat Genet*. 2015 Feb 23.
12. Calon A, Lonardo E, Berenguer-Llergo A, Espinet E, Hernando-Momblona X, Iglesias M, et al. Stromal gene expression defines poor-prognosis subtypes in colorectal cancer. *Nat Genet*. 2015 Feb 23.

13. Li X, Madison BB, Zacharias W, Kolterud A, States D, Gumucio DL. Deconvoluting the intestine: molecular evidence for a major role of the mesenchyme in the modulation of signaling cross talk. *Physiol Genomics*. 2007 May 11;29(3):290-301.
14. Jorissen RN, Gibbs P, Christie M, Prakash S, Lipton L, Desai J, et al. Metastasis-Associated Gene Expression Changes Predict Poor Outcomes in Patients with Dukes Stage B and C Colorectal Cancer. *Clin Cancer Res*. 2009 Dec 15;15(24):7642-51.
15. Ong CW, Chong PY, McArt DG, Chan JY, Tan HT, Kumar AP, et al. The prognostic value of the stem-like group in colorectal cancer using a panel of immunohistochemistry markers. *Oncotarget*. 2015 May 20;6(14):12763-73.
16. Dunne PD, McArt DG, Blayney JK, Kalimutho M, Greer S, Wang T, et al. AXL is a key regulator of inherent and chemotherapy-induced invasion and predicts a poor clinical outcome in early-stage colon cancer. *Clin Cancer Res*. 2014 Jan 1;20(1):164-75.
17. Dunne PD, Dasgupta S, Blayney J, McArt DG, Redmond KL, Weir JA, et al. EphA2 expression is a key driver of migration and invasion and a poor prognostic marker in colorectal cancer. *Clin Cancer Res*. 2015 Aug 17.
18. Karamitopoulou E, Zlobec I, Koelzer VH, Langer R, Dawson H, Lugli A. Tumour Border Configuration in Colorectal Cancer: Proposal for an Alternative Scoring System based on the Percentage of Infiltrating Margin. *Histopathology*. 2015 Feb 4.
19. Huijbers A, Tollenaar RA, v Pelt GW, Zeestraten EC, Dutton S, McConkey CC, et al. The proportion of tumor-stroma as a strong prognosticator for stage II and III colon cancer patients: validation in the VICTOR trial. *Ann Oncol*. 2013 Jan;24(1):179-85.
20. Berdiel-Acer M, Berenguer A, Sanz-Pamplona R, Cuadras D, Sanjuan X, Paules MJ, et al. A 5-gene classifier from the carcinoma-associated fibroblast transcriptomic profile and clinical outcome in colorectal cancer. *Oncotarget*. 2014 Aug 15;5(15):6437-52.
21. Jass JR. Classification of colorectal cancer based on correlation of clinical, morphological and molecular features. *Histopathology*. 2007 Jan;50(1):113-30.
22. Gerlinger M, Rowan AJ, Horswell S, Larkin J, Endesfelder D, Gronroos E, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med*. 2012 Mar 8;366(10):883-92.
23. Kim TM, Jung SH, An CH, Lee SH, Baek IP, Kim MS, et al. Subclonal Genomic Architectures of Primary and Metastatic Colorectal Cancer Based on Intratumoral Genetic Heterogeneity. *Clin Cancer Res*. 2015 Oct 1;21(19):4461-72.
24. Rodon J, Soria JC, Berger R, Batist G, Tsimberidou A, Bresson C, et al. Challenges in initiating and conducting personalized cancer therapy trials:

perspectives from WINTHER, a Worldwide Innovative Network (WIN) Consortium trial. *Ann Oncol*. 2015 Aug;26(8):1791-8.

25. Lazar V, Rubin E, Depil S, Pawitan Y, Martini JF, Gomez-Navarro J, et al. A simplified interventional mapping system (SIMS) for the selection of combinations of targeted treatments in non-small cell lung cancer. *Oncotarget*. 2015 Jun 10;6(16):14139-52.

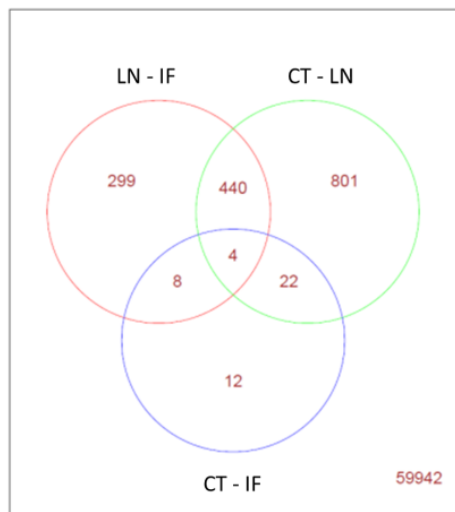
26. Westin JR, Kurzrock R. It's about time: lessons for solid tumors from chronic myelogenous leukemia therapy. *Mol Cancer Ther*. 2012 Dec;11(12):2549-55.

27. Salto-Tellez M, James JA, Hamilton PW. Molecular pathology - the value of an integrative approach. *Mol Oncol*. 2014 Oct;8(7):1163-8.

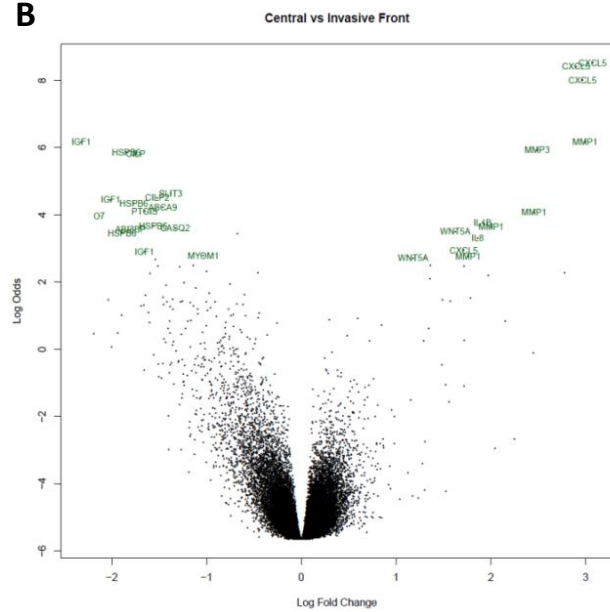
28. McArt DG, Blayney JK, Boyle DP, Irwin GW, Moran M, Hutchinson RA, et al. PICan: An integromics framework for dynamic cancer biomarker discovery. *Mol Oncol*. 2015 Jun;9(6):1234-40.

Figure 1

A



B



C

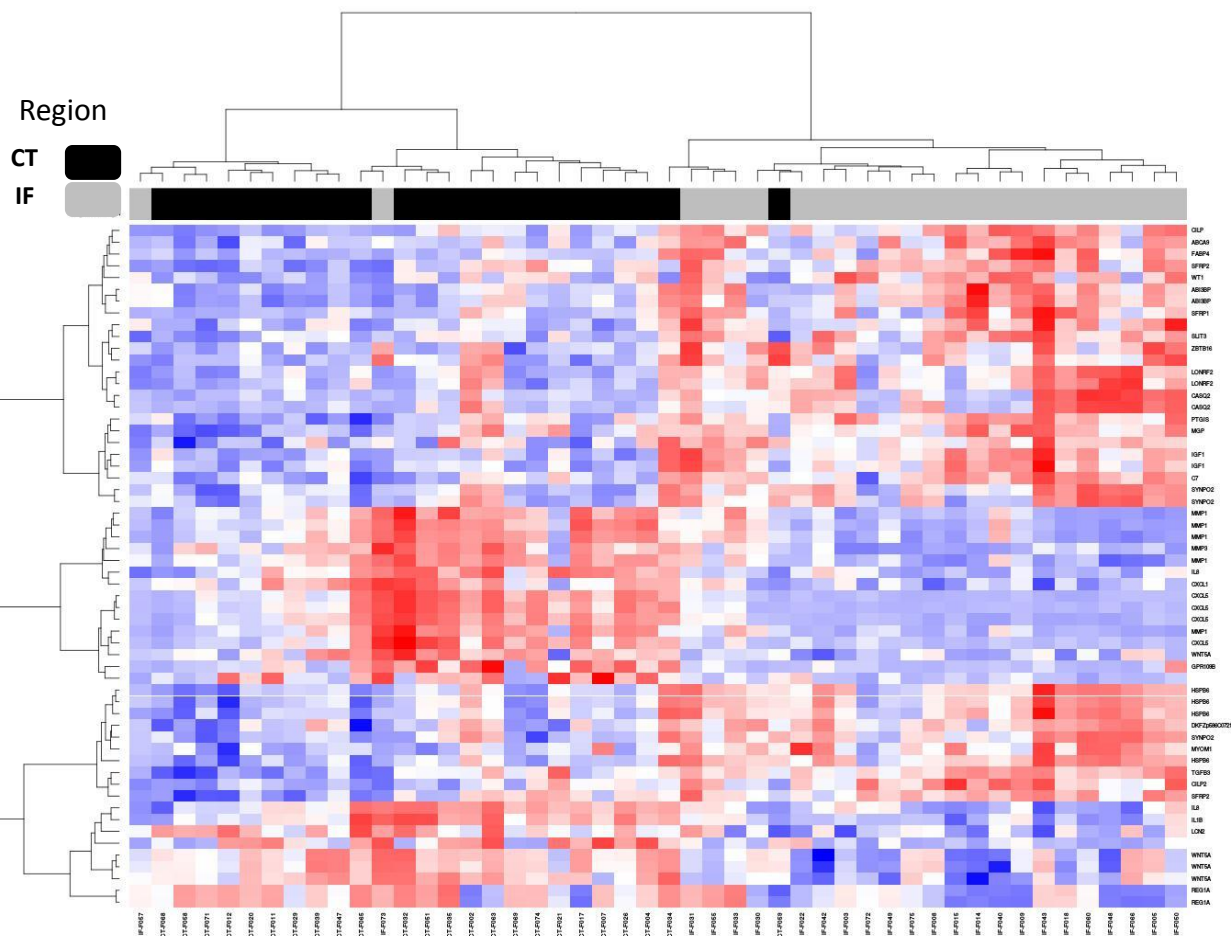


Figure 3

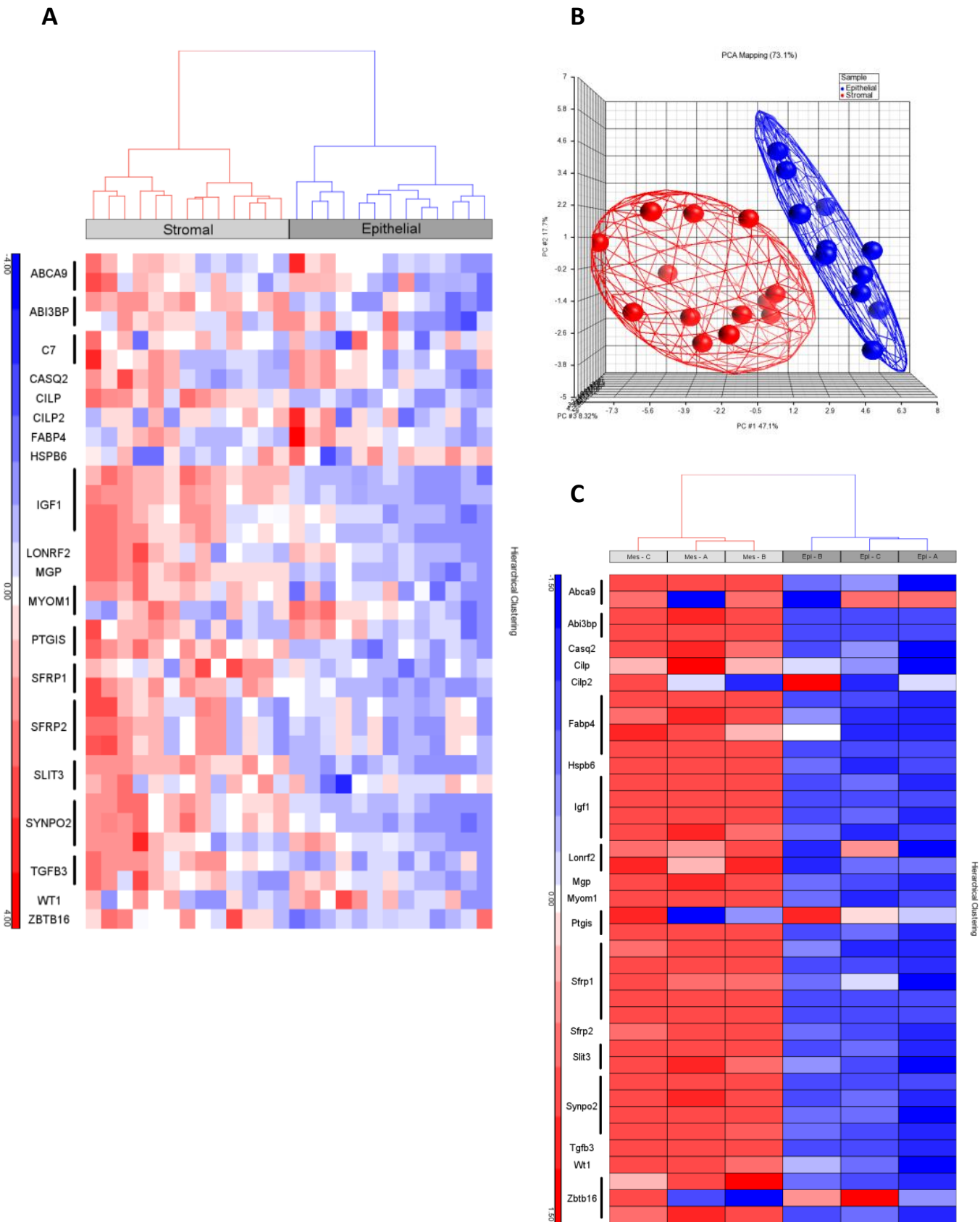


Figure 4

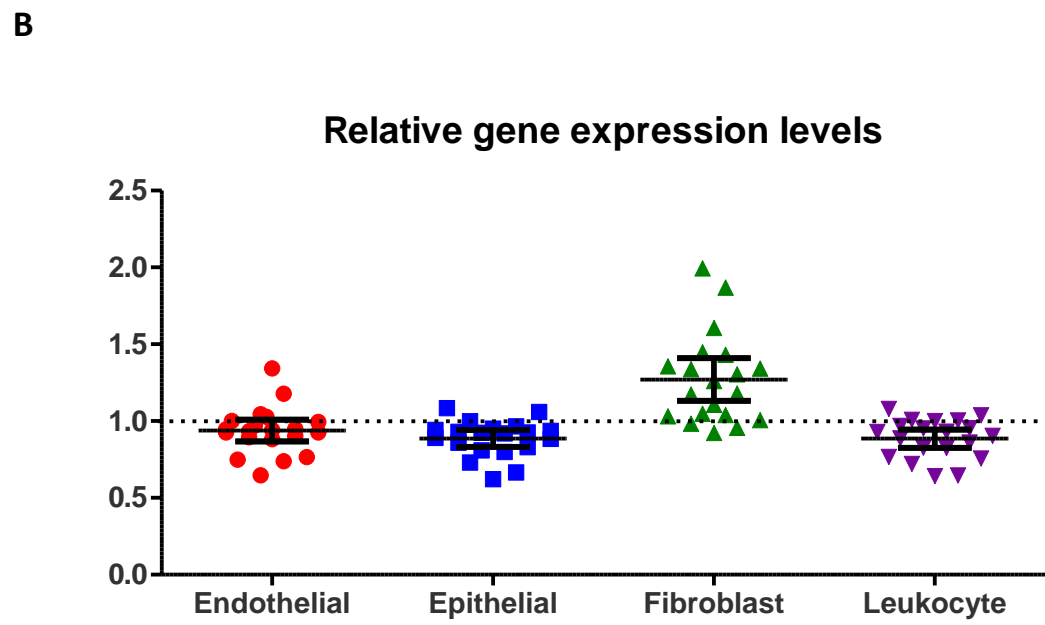
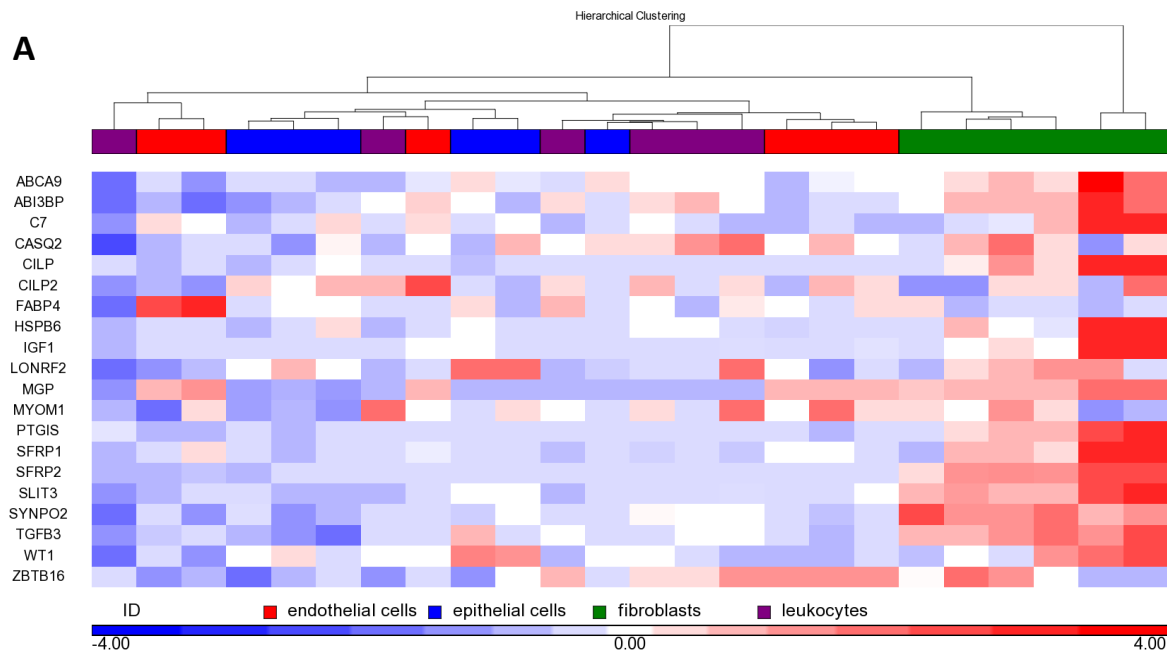
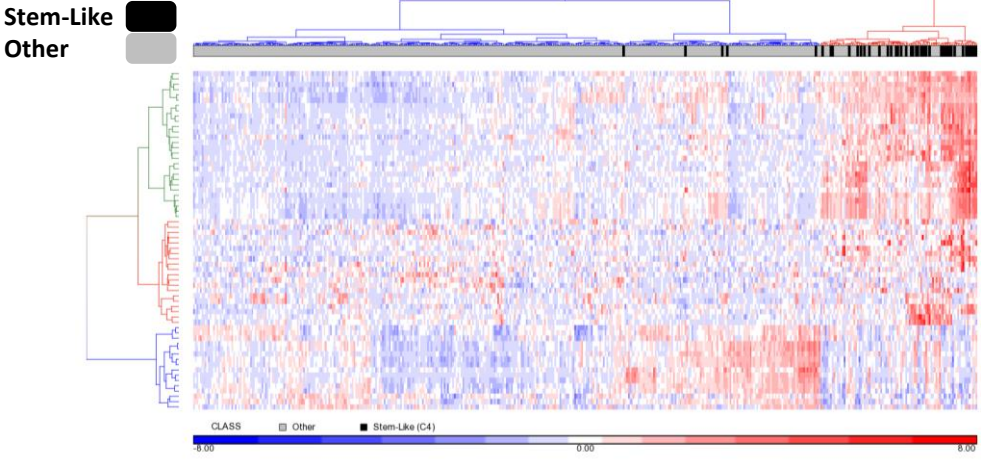
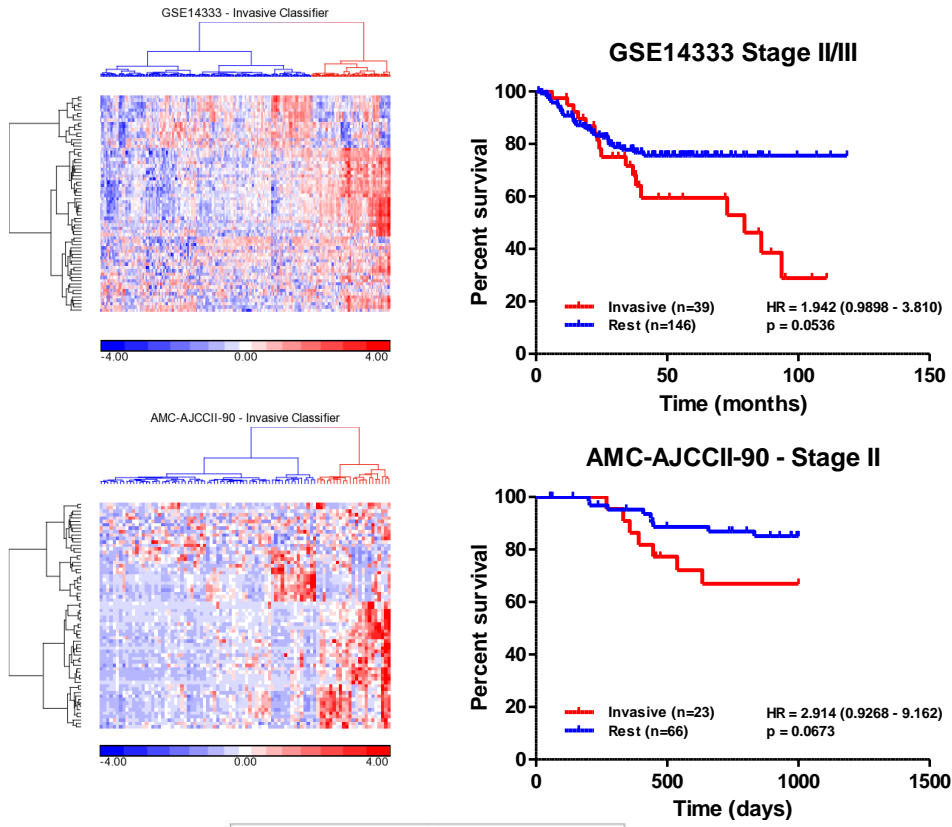


Figure 5

A



B



C

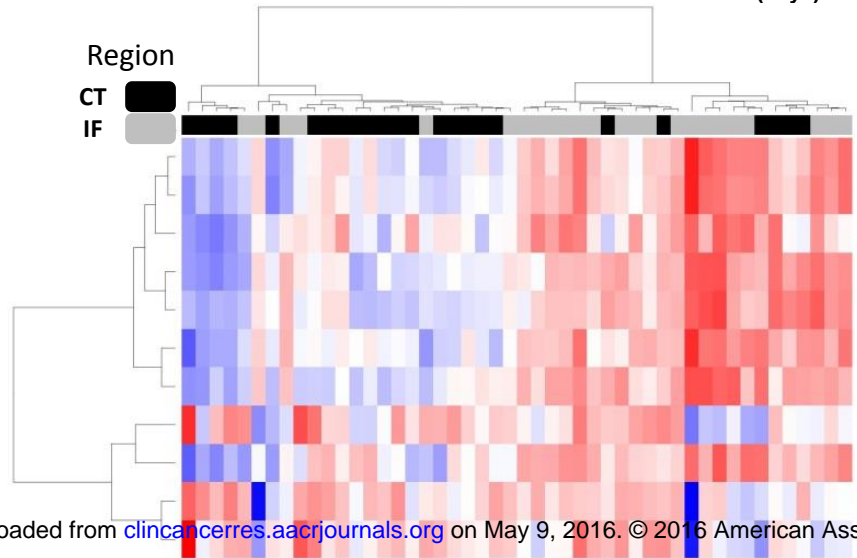


Table 1

Symbol	Entrez Gene Name
ABCA9	ATP-binding cassette, sub-family A (ABC1), member 9
ABI3BP	ABI family, member 3 (NESH) binding protein
C7	complement component 7
CASQ2	calsequestrin 2 (cardiac muscle)
CILP	cartilage intermediate layer protein
CILP2	cartilage intermediate layer protein 2
FABP4	fatty acid binding protein 4, adipocyte
HSPB6	heat shock protein, alpha-crystallin-related, B6
IGF1	insulin-like growth factor 1 (somatomedin C)
LONRF2	LON peptidase N-terminal domain and ring finger 2
MGP	matrix Gla protein
MYOM1	myomesin 1
PTGIS	prostaglandin I2 (prostacyclin) synthase
SFRP1	secreted frizzled-related protein 1
SFRP2	secreted frizzled-related protein 2
SLIT3	slit homolog 3
SYNPO2	synaptopodin 2
TGFB3	transforming growth factor, beta 3
WT1	Wilms tumor 1
ZBTB16	zinc finger and BTB domain containing 16
CXCL1	chemokine (C-X-C motif) ligand 1
CXCL5	chemokine (C-X-C motif) ligand 5
CXCL8	chemokine (C-X-C motif) ligand 8
HCAR3	hydroxycarboxylic acid receptor 3
IL1B	interleukin 1, beta
MMP1	matrix metalloproteinase 1
MMP3	matrix metalloproteinase 3
REG1A	regenerating islet-derived 1 alpha
WNT5A	wingless-type MMTV integration site family, member 5A
LCN1	lipocalin 1

Clinical Cancer Research

Challenging the cancer molecular stratification dogma: Intratumoral heterogeneity undermines consensus molecular subtypes and potential diagnostic value in colorectal cancer

Philip D. Dunne, Darragh G McArt, Conor A. Bradley, et al.

Clin Cancer Res Published OnlineFirst May 5, 2016.

Updated version	Access the most recent version of this article at: doi: 10.1158/1078-0432.CCR-16-0032
Supplementary Material	Access the most recent supplemental material at: http://clincancerres.aacrjournals.org/content/suppl/2016/05/05/1078-0432.CCR-16-0032.DC1.html
Author Manuscript	Author manuscripts have been peer reviewed and accepted for publication but have not yet been edited.

E-mail alerts	Sign up to receive free email-alerts related to this article or journal.
Reprints and Subscriptions	To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org .
Permissions	To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org .